ALIGNING DATA STREAMS

Michele M. Covell

Harold G. Sampson

BACKGROUND OF THE INVENTION

5 1.   Field of the Invention

This invention relates to aligning data streams (e.g.,
sets of visual and/or audio data). The invention
particularly relates to quantized alignment (i.e., alignment
at a lower temporal resolution than that of the data that is
10 used to do the alignment) of wide-bandwidth data streams.
The invention also particularly relates to selecting
distinctive audio segments for cross-correlation to enable
alignment of data streams including audio data.

2.   Related Art

15   There are various situations in which it is desirable to
use high resolution information to provide quantized
estimates of the optimal alignment between two data streams.
An example of this is using audio samples from two sets of
audiovisual data to estimate how many video frames (which are
20 obtained at a relatively low rate compared to that at which
audio samples are obtained) to offset one video stream
relative to another for optimal alignment of the audio and,
by association, the video frames and (if applicable) the
associated metadata of the two sets of audiovisual data. In
25 such case, since it does not make sense, from the video point
of view, to talk about offsets other than in video frame rate
increments, a situation exists in which the data that it is
desired to use for alignment (the audio data) is much higher
resolution than the alignment information that it is desired
30 to estimate.

An approach could be taken of cross-correlating the two
data streams at the high resolution of the data (e.g., at the
resolution of audio samples) and then, after finding the
highest normalized correlation location, quantizing that

location to a lower resolution of the data (e.g., to a
multiple of the video frame rate). This approach has the
disadvantage of requiring more computation than would
nominally be expected for the number of distinct alignment
5 possibilities that are ultimately being considered.

Another approach would be to use the high resolution
data (e.g., audio samples), but only sample the cross-
correlation at a lower resolution (e.g., once every video
frame period). This has the distinct disadvantage of
10 undersampling the cross-correlation function relative to its
Nyquist rate: since the cross-correlation function is not
being sampled often enough, it is very likely that the
optimal alignment will be missed and, instead, some other
alignment selected that is far from the best choice. See
15 A.V. Oppenheim, R.W. Schafer, Discrete-Time Signal Processing
(Prentice Hall, 1989), for more detailed discussion of
undersampled signals and aliasing.

Still another approach would be to low pass (or band
pass) filter the high resolution data streams before
20 attempting the cross-correlation. In this case, the cross-
correlation function can be sampled at the lower resolution
without worrying about the Nyquist rate: the low pass (or
band pass) filter of the inputs into the cross-correlation
function ensures that the Nyquist requirements are met.
25 However, low pass (or band pass) filtering the input data so
severely is likely to remove many of the distinctive
identifying characteristics of the high resolution data
streams, thus degrading the ability of the cross-correlation
to produce accurate alignment. For example, if this approach
30 is used with two audiovisual data streams, even if a "good"
band is selected to pass, there are not many distinguishing
features left in an audio signal that has been filtered down
to a 15 Hz bandwidth (15Hz = 30Hz/2), since sampling occurs
at 30 Hz and Nyquist requires 2 samples/cycle.

Additionally, there are various situations in which it is desired to use a short segment from each of two long audio data streams to estimate an alignment between the two audio data streams and any associated data (e.g., video data,

5 metadata). An example of this is using audio samples from two sets of audiovisual data to estimate how many frames to offset one video stream relative to another for optimal alignment of the audio and, by association, the video frames and (if applicable) the associated metadata of the two sets

10 of audiovisual data. Since the amount of computation that is required for the cross-correlation varies as N log N, where N is the segment length that is being used in the cross-correlation, it is typically not desirable to use the full audio streams. Instead, it is desirable to select a short

15 segment from one of the audio streams that is both stable (i.e., unlikely to "look different" after repeated digitization) and distinctive. (Stability can be an issue, for example, in applications in which a first digitization uses automatic gain control and a second digitization

20 doesn't, so that it is necessary to be careful about picking segments with low power in the frequency bands at which the automatic gain control responds.) If these two criteria are met, a single, clear-cut correlation peak that is well localized and is well above the noise floor can be obtained.

25 One way to select such a short segment would be to examine the auto correlation function over local windows. This approach has the disadvantage of being computationally expensive: it requires on the order of N log N computations for each N-length local window that is considered.

30 SUMMARY OF THE INVENTION

According to one aspect of the invention, two wide-bandwidth, high resolution data streams are aligned, in a manner that retains the full bandwidth of the data streams, by using magnitude-only spectrograms as inputs into the

cross-correlation and sampling the cross-correlation at a
coarse sampling rate that is the final alignment quantization
period.

5 According to another aspect of the invention, stable and
distinctive audio segments are selected for cross-correlation
by evaluating the energy in local audio segments and the
variance in energy among nearby audio segments.

DETAILED DESCRIPTION OF THE INVENTION

I.   Quantized Alignment of Wide-Bandwidth Data Streams
10   According to one aspect of the invention, wide-
bandwidth, high resolution data streams can be aligned at a
lower resolution in a manner that retains the full bandwidth
of the data, but only samples the cross-correlation at a
coarse sampling rate that is a final alignment quantization
15 period corresponding to the lower resolution. For example,
this aspect of the invention can be used to align audiovisual
data streams at the resolution of the video data, using the
audio data to produce the alignment.

Quantized alignment of wide-bandwidth data streams
20 according to this aspect of the invention avoids problems
with undersampling by using magnitude-only spectrograms as
inputs into the cross-correlation. A magnitude-only
spectrogram is computed for each of the high resolution data
streams, using a spectrogram slice length (e.g., video frame
25 size) that is appropriate for the stationarity
characteristics of the high resolution data streams (i.e.,
that produces largely stationary slices of the high
resolution data) and a spectrogram step size (e.g., video
frame offset) that is appropriate for the quantization period
30 of the final alignment (i.e., that can achieve the resolution
requirements of the low-resolution alignment). If the
spectrogram slices are too short, the spectrogram slices can
suffer from strong local edge effects (e.g., in audio,
glottal pulses). On the other hand, it is desirable for the

spectrogram slices to be no longer than the desired
resolution of the alignment. However, the latter
consideration is less important than the former; if there is
a conflict between the two, the former consideration should
5 govern the selection of spectrogram slice length. When this
aspect of the invention is used to align audiovisual data
streams, the spectrogram slice length and step size can be,
for example, 1/29.97 sec. (which corresponds to a common
video frame rate).

10   Treating the spectrograms as multi-channel data vectors,
one-dimensional cross-correlation can be used on these low
sampling rate streams. For example, any of the standard FFT-
based one-dimensional convolution routines can be used.

      Quantized alignment of wide-bandwidth data streams
15 according to this aspect of the invention reduces overall
computational requirements compared to previous approaches.
For example, for the first approach described in the
Background section above, with the minimum-sized FFT-based
cross-correlation being used for computational efficiency
20 (i.e., allowing aliasing on the offsets that will not be
examined), the computational load is $3/2 * (N+L) * M *
(\log(N+L) + \log M)$, where N is the maximum forward or
backward offset that it is desired to consider (that is, +/-
N), M is the oversampling rate of the high resolution data
25 stream, and L*M is the number of samples over which
integration will be performed to get a cross-correlation
estimate. In contrast, for the invention, the computational
load is $L * M * \log(M)$ for the spectrograms, plus $3/2 * M *
(N+L) * \log(N+L)$ for the M-channel, low-resolution cross-
30 correlation. The computational savings, as compared to the
first approach described in the Background section above, is
$[3/2 * M * N * \log(M)] + [1/2 * M * L * \log(M)]$. For some
typical audio/video settings, M=500, N=20, and L=160. In
that case, there is a 22% computational savings. The
35 reduction in memory requirements is much larger: the size of

the individual FFTs that are used are reduced by a factor of M (in a typical audio/video setting, 500), resulting in a similar reduction in fast memory requirements.

II.   Selecting Distinctive Audio Segments for Cross-
5   Correlation

According to another aspect of the invention, a conservative approach is used to select distinctive audio segments for cross-correlation, which will tend to err on the side of not finding a segment that could have been accurately
10 used for cross-correlation and will seldom return a segment that is not reasonable for finding a good cross-correlation peak.

According to this aspect of the invention, distinctive audio segments for cross-correlation are selected using an
15 approach that is based on the energy in a local audio segment and how the energy varies in nearby audio segments. In one implementation of this aspect of the invention, energy measures are computed using three window lengths: a short time window (e.g., 0.125 sec.) for computing local audio
20 energy, a long time window (e.g., the whole audio stream) for computing normalizing constants, and a mid-length time window (e.g., 1 sec.) for computing the local variation in the audio energy level.

According to this aspect of the invention, segments are
25 marked as "good segments" (i.e., segments that can be used for cross-correlation) when the audio energy level for the segment is above some minimum threshold (e.g., 0.3 times the global mean energy, A_{var}) and when the audio energy level in nearby segments varies by some other minimum threshold (e.g.,
30 the variance of the audio energy over the mid-length time window varies by at least 0.1 times the square of the global mean energy, A_{var}).

This aspect of the invention is desirably further implemented to ensure that random noise segments are not

selected when the audio stream is essentially silent. This
aspect of the invention can be implemented to avoid that
problem by adjusting the estimate of the global mean energy,
A_{var}, upward, whenever the estimate of the global mean energy,
5 A_{var}, is less than the square of the global mean level.

In summary, if the audio stream is x[n], the global mean
energy, A_{var}, is established according to the following
equation:

$$10 \qquad A_{var} = MAX \left(\left(1/T \sum_{N=0}^{T-1} m_R\{x,N\} \right)^2, \ 1/T \sum_{N=0}^{T-1} v_R\{x,N\} \right) \qquad (1)$$

$$\text{where } m_R\{x,N\} = 1/R \sum_{n=0}^{R-1} x[n+RN] \qquad (2)$$

15

$$v_R\{x,N\} = 1/R \sum_{n=0}^{R-1} x^2[n+RN] - m_R^2\{x,N\} \qquad (3)$$

    R = length of the short time window
    RT = length of the long time window
20   T = constant relating the length of the long time
            window to the length of the short time window

The audio segments on which to cross-correlate are selected
from the set of segments that satisfy the follow conditions
(the outer local mean and outer local variance estimates are
25 taken using "N" as the sequence index):

$$m_S\{v_R(x,N),k\} > T_{level}A_{var} \qquad (4)$$
$$v_S\{v_R(x,N),k\} > T_{var}A^2_{var} \qquad (5)$$

where RS = length of the mid-length time window
       S = constant relating length of the mid-length time
30   window to length of the short time window
     T_{level} = constant establishing threshold audio energy
          level for audio segment to be identified as
          distinctive
     T_{var} = constant establishing threshold audio energy
35   variance for audio segment to be identified as
          distinctive

The length R of the short time window can be established
as some constant multiple (e.g., 1) of the low-resolution
alignment that it is desired to achieve. The constant S can
be chosen so that the length of the mid-length time window is
5 short enough to achieve a desired computational efficiency
and long enough to be effective in disambiguating multiple
correlation peaks. A particular value of S can be chosen
empirically in view of the above-described considerations.
Unless prohibitively computationally expensive, the
10 constant T can be chosen so that the long time window is
equal to the duration of the entire set of audio data from
which the audio segments are to be chosen for cross-
correlation.

Various embodiments of the invention have been
15 described. The descriptions are intended to be illustrative,
not limitative. Thus, it will be apparent to one skilled in
the art that certain modifications may be made to the
invention as described herein without departing from the
scope of the claims set out below.